

Белорусский государственный университет

УТВЕРЖДАЮ

Проректор по учебной работе



А.И. Толстик

25 июля 2016 г.



Регистрационный № УД 3107 уч.

ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

**Учебная программа учреждения высшего образования
по учебной дисциплине для специальности:**

1-31 03 07 Прикладная информатика (по направлениям)

Направление специальности:

1-31 03 07-02 Прикладная информатика (информационные
технологии телекоммуникационных систем)

Учебная программа составлена на основе ОСВО 1-31 03 07-2013 и типовых учебных планов УВО №G31-170/уч. 2013 г., УВО №G31-186/уч. 2013 г.

СОСТАВИТЕЛИ:

Николай Николаевич Яцков, доцент кафедры системного анализа и компьютерного моделирования, кандидат физико-математических наук, доцент

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой системного анализа и компьютерного моделирования Белорусского государственного университета (протокол № 12 от 14.06.2016 г.);

Учебно-методической комиссией факультета радиофизики и компьютерных технологий Белорусского государственного университета (протокол № 10 от 21.06.2016 г.)

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Учебная программа дисциплины специализации «Введение в биоинформатику» разработана для студентов специальности 1-31 03 07 Прикладная информатика (по направлениям), направление специальности 1-31 03 07-02 Прикладная информатика (информационные технологии телекоммуникационных систем), специализации 1-31 03 07-02-04 Биоинформатика в соответствии с требованиями ОСВО по специальности 1-31 03 07 «Прикладная информатика» ОСВО 1-31 03 07 и учебного плана БГУ по специальности 1-31 03 07 «Прикладная информатика», направление специальности 1-31 03 07-02 «Прикладная информатика (информационные технологии телекоммуникационных систем)».

Целью учебной дисциплины является изучение теоретических основ биоинформатики, включая базовые элементы молекулярной биологии, методы регистрации и анализа биологических последовательностей, основы статистического программирования и интеллектуального анализа данных с использованием языка R.

Задачи дисциплины – научить производить расчеты с применением методов биоинформатики и решать широкий спектр прикладных задач обработки биологических наборов данных в среде статистического программирования R.

Учебная программа составлена с учетом межпредметных связей и программ по учебным дисциплинам «Теория вероятностей и математическая статистика», «Программирование», «Интеллектуальный анализ данных».

Изучение учебной дисциплины «Введение в биоинформатику» должно обеспечить формирование у студента следующих компетенций:

АК-1. Уметь применять базовые научно-теоретические знания для решения теоретических и практических задач.

АК-4. Уметь работать самостоятельно.

АК-6. Владеть междисциплинарным подходом при решении проблем.

АК-7. Иметь навыки, связанные с использованием технических устройств, управлением информацией и работой с компьютером.

ПК-4. Разрабатывать численные алгоритмы и программы.

ПК-7. Применять профессиональные знания и навыки для проведения научных исследований в области прикладной информатики.

ПК-8. Разрабатывать и совершенствовать методы исследований в области информационных и телекоммуникационных систем.

ПК-9. Работать с научно-технической информацией с использованием современных информационных технологий.

ПК-10. Формулировать выводы и рекомендации по применению результатов научно-исследовательской работы.

ПК-11. Пользоваться глобальными информационными ресурсами.

В результате изучения дисциплины студент должен:

знать:

- базовые понятия и принципы биоинформатики,
- основные алгоритмы методов и моделей биоинформатики и подходы к их созданию,

- задачи анализа больших наборов биологических данных;

уметь:

- производить расчеты с применением алгоритмов биоинформатики,
- применять методы биоинформатики в среде статистического программирования R для решения практических задач управления и обработки биоинформации,
- творчески и эффективно использовать полученные знания в профессиональной деятельности.

В соответствии с учебными планами объем дисциплины составляет 144 учебных часа, в том числе 66 аудиторных часов, из них лекции – 34, лабораторные работы – 32.

Дисциплина «Введение в биоинформатику» изучается студентами дневной формы получения высшего образования на 4-ом курсе в 7 семестре.

Форма текущей аттестации по учебной дисциплине – зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Введение в биоинформатику.

1.1. Основные понятия дисциплины «Введение в биоинформатику». Структура биоинформатики. Информационные ресурсы. Задачи, методы и модели биоинформатики. Данные и базы данных. Понятие данных. Форматы данных. Геномные и белковые базы данных. Примеры практического применения биоинформатики. Фармацевтика. Медицина. Пищевая промышленность. Экология. Аэрокосмические исследования.

Раздел 2. Биоинформационные системы.

2.1. Биологические последовательности и центральная догма молекулярной биологии. Биология клетки. Биологические последовательности. Центральная догма молекулярной биологии. Ген. Генетический код. Транскрипция. Экзоны. Интроны. Сплайсинг. Альтернативный сплайсинг. Трансляция. Экспрессия гена. Транскрипт. Геном. Выравнивание биологических последовательностей. Глобальное и локальное выравнивание. Парное и множественное выравнивание биологических последовательностей.

2.2. Алгоритмы сравнения биологических последовательностей. Алгоритмы сравнения с использованием выравнивания. Точечная матрица сходства. Расстояния Хэмминга и Левенштейна. Алгоритм динамического программирования. Алгоритм Нидлмана-Вунша. Алгоритм Смита-Уотермана. Алгоритмы FASTA и BLAST. Алгоритмы сравнения без выравнивания. Алгоритм k-tupels. Алгоритм CPF.

2.3. Технологии секвенирования ДНК. Технологии секвенирования ДНК. Основные принципы классического секвенирования ДНК. Секвенирование ДНК. Геномные библиотеки. Риды. Гель-электрофорез. Полимеразная цепная реакция. Секвенирование по Максому-Гилберту. Секвенирование по Сангеру. Технологии геномного секвенирования NGS. Roche 454. Ion Torrent. Illumina. Pacific Biosciences. Nanopore. Проект Bioconductor.

Раздел 3. Биостатистическое программирование на языке R.

3.1. История и основные принципы организации среды R. Программные ресурсы для анализа данных. Типы данных. Константы. Атрибуты объектов. Базовые операторы и функции. Векторы. Скрытая коэрсия. Матрицы. Списки. Факторы. Пропущенные значения. Таблицы. Имена.

3.2. Работа с данными в R. Управление наборами данных. Доступ к элементам структур данных, матриц и списков. Частичное совпадение (Partial matching). Векторные операции. Чтение и запись данных. Функции read.table, readLines, source, dget, load, write.table, writeLines, dump, dput.

3.3. Организация вычислений в R. Управляющие структуры if, else, for, while, repeat, break, next, return. Функции. Сопоставление аргументов функций. Ленивое вычисление (Lazy Evaluation). Правила следования аргументов функций. Правила видимости свободных переменных в функциях.

3.4. Обработка даты и времени, векторизованные вычисления. Дата и время. Векторизованные функции. Функции lapply, sapply, apply, tapply, mapply, split.

3.5. Отладка программного кода и имитационное моделирование.

Поиск и исправление ошибок в программном коде. Сообщения message, warning и error. Функции отладки программного кода traceback, debug, browser, trace, recover, print. Программная оболочка для среды R – RStudio. Имитационное моделирование. Базовые функции R для моделирования дискретных и непрерывных случайных величин. Моделирование непрерывной случайной величины методами обратных функций и Неймана. Моделирование дискретной случайной величины. Моделирование нуклеотидной последовательности. Формирование случайной выборки.

Раздел 4. Интеллектуальный анализ биологических данных в среде R.

4.1. Предварительный анализ данных. Описательная статистика. Характеристики центральной тенденции. Характеристики вариации. Графическое представление данных в R. Двумерный график. Гистограмма. Изоповерхности и контурные линии. Коробчатая диаграмма. Столбиковые диаграммы. Диаграмма рассеяния. Диаграмма рассеяния в 3D. Поверхность функции. Очистка данных. Нормировка и стандартизация данных. Анализ выбросов и аномальных значений.

4.2. Корреляционный и регрессионный анализ. Корреляционный анализ. Коэффициент корреляции Пирсона. Определение значимости коэффициента корреляции. Ранговая корреляция. Критерий Спирмена. Критерий Кендэла. Частная корреляция. Регрессионный анализ. Общие модели линейной и нелинейной регрессий. Проверка точности регрессионной модели.

4.3. Дисперсионный анализ и методы снижения размерности данных. Дисперсионный анализ. Однофакторный и двухфакторный дисперсионный анализ. Методы снижения размерности данных. Метод главных компонент. Факторный анализ.

4.4. Иерархические методы кластерного анализа. Кластерный анализ. Основные элементы кластерного анализа. Этапы кластерного анализа. Типы кластеров. Расстояния между объектами данных. Математические характеристики кластера. Критерии качества кластеризации. Классификация алгоритмов кластерного анализа. Иерархические методы кластерного анализа. Дендрограмма. Формула Ланса-Уильямса. Расстояния между кластерами данных. Иерархический агломеративный, дивизимный и гибридный кластерный анализ.

4.5. Неиерархические методы кластерного анализа. Алгоритмы на основе k-средних. Алгоритм k-средних. Алгоритм Fuzzy k-средних. Алгоритмы k-медоид. PAM, PAMk, CLARA, DBSCAN.

4.6. Методы классификации: методы k-ближайших соседей и байесовских сетей. Методы классификации данных. Алгоритмы k-ближайших соседей. Байесовская классификация.

4.7. Методы классификации: методы деревьев решений и нейронных сетей. Деревья решений. Анализ данных с использованием деревьев решений. Методика «разделяй и властвуй». Критерии и функции качества разбиения узлов дерева. Индекс Джини. Энтропия. Ошибка классификации. Остановка построения дерева. Сокращение дерева. Обработка пропущенных значений.

Извлечение правил из деревьев. Алгоритмы построения деревьев решений. Алгоритмы Conditional Inference Tree, CART, Random Forests.

4.8. Методы поиска ассоциативных правил и визуализация многомерных данных. Нейронные сети. Нейрон. Нейронная сеть. Нейронные сети Кохонена. Нейронная сеть слой Кохонена. Самоорганизующиеся карты Кохонена. Методы поиска ассоциативных правил. Ассоциативные правила. Алгоритм Apriori. Визуальное представление данных. Ресурсы R для визуального представления данных.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов		Количество часов УСР	Форма контроля знаний
		Лекции	Лабораторные занятия		
1	Введение в биоинформатику (2 ч)	2			
1.1	Основные понятия дисциплины «Введение в биоинформатику»				
2	Биоинформационные системы (6 ч)	6	8		
2.1	Биологические последовательности и центральная догма молекулярной биологии	2			Тест
2.2	Алгоритмы сравнения биологических последовательностей	2	8		Отчет по лабораторной работе
2.3	Технологии секвенирования ДНК	2			Тест
3	Биостатистическое программирование на языке R (6 ч)	10	8		
3.1	История и основные принципы организации среды R	2	8		Отчет по лабораторной работе
3.2	Работа с данными в R	2			Тест
3.3	Организация вычислений в R	2			
3.4	Обработка даты и времени, векторизированные вычисления	2			Тест
3.5	Отладка программного кода и имитационное моделирование	2			
4	Интеллектуальный анализ биологических данных в среде R (16 ч)	16	16		
4.1	Предварительный анализ данных	2	4		Отчет по лабораторной работе
4.2	Корреляционный и регрессионный анализ	2			
4.3	Дисперсионный анализ и методы снижения размерности данных	2	4		Отчет по лабораторной работе
4.4	Иерархические методы кластерного анализа	2	4		Отчет по лабораторной работе
4.5	Неиерархические методы кластерного анализа	2	4		Отчет по лабораторной работе
4.6	Методы классификации: методы k-ближайших соседей и байесовских сетей	2			Тест
4.7	Методы классификации: методы деревьев решений и нейронных сетей	2			
4.8	Методы поиска ассоциативных правил и визуализация многомерных данных	2			Тест

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Основная и дополнительная литература

№№ п/п	Список литературы
	Основная
1.	Введение в биоинформатику / А.Леск; пер. с англ. – М.: БИНОМ. Лаборатория знаний, 2009. – 318 с.
2.	Основы биоинформатики : учеб. Пособие / А.Н. Огурцов. - Х.: НТУ «ХПИ», 2013. – 400 с.
3.	Bioinformatics: Sequence and Genome Analysis / David W. Mount, Gold Spring Harbor Laboratoty Press. 2004 - 565 pp.
4.	Мастицкий С.Э., Шитиков В.К. Статистический анализ и визуализация данных с помощью R / Хайдельберг – Лондон – Тольятти. 2014. - 401 с.
5.	R в действии. Анализ и визуализация данных в программе R / Кабаков Р. И. пер. с англ. – М.: ДМК Пресс, 2014. – 588 с.
6.	Яцков, Н. Н. Интеллектуальный анализ данных : пособие / Н. Н. Яцков. – Минск : БГУ, 2014. – 151 с.
	Дополнительная
1.	Learning Data Mining with R / Batek Makhlal. Packt Publishing. 2015. 314 pp.
2.	R and Data Mining: Examples and Case Studies / Yanchang Zhao. Elsevier. 2012. 164 pp.
3.	Дюк, В. Data Mining: учебный курс / В. Дюк, А. Самойленко. – СПб., 2001. 368 с.
4.	R in a nutshell / Joseph Adler. Second Edition. O'Reilly Media, Inc. 2012. 722 pp.
5.	Ивченко, Г. И. Введение в математическую статистику : учебник / Г. И. Ивченко, Ю. И. Медведев. – М., 2010. – 600 с.
6.	Кобзарь, А. И. Прикладная математическая статистика : для инженеров и научных работников / А. И. Кобзарь. – М., 2006. – 816 с.
7.	Мандель, И.Д. Кластерный анализ / И.Д. Мандель –М., 1988. – 176 с.
8.	Data Mining with R. Learning with Case Studies / Luís Torgo. Chapman & Hall/CRC. 2011. 289 pp.
9.	Наследов, А. Д. Математические методы психологического исследования. Анализ и интерпретация данных : учеб. пособие / А. Д. Наследов. – СПб., 2004. – 392 с.
10.	Сегаран, Т. Программируем коллективный разум : пер. с англ. / Т. Сегаран. – СПб., 2008. – 368с.

Перечень тем лабораторных работ:

1. Основы работы в R (8 часов).
2. Предварительный анализ больших наборов биоданных (4 часа).
3. Метод главных компонент (4 часа).
4. Иерархические методы кластерного анализа (4 часа).
5. Неиерархические методы кластерного анализа (4 часа).
6. Алгоритмы сравнения биологических последовательностей (8 часов).

Методические рекомендации по организации самостоятельной работы студентов

Для организации самостоятельной работы студентов по учебной дисциплине рекомендуется использовать современные информационные технологии: разместить в сетевом доступе комплекс учебных и учебно-методических материалов (программа курса, методические указания к лабораторным занятиям, задания в тестовой форме, список рекомендуемой литературы и информационных ресурсов и др.).

Перечень используемых средств диагностики результатов учебной деятельности студентов

В качестве формы итогового контроля по дисциплине используется зачет. Оценка учебных достижений студента на зачете производится по десятибалльной шкале.

Для оценки уровня знаний студентов используется следующий диагностический инструментарий:

- компьютерные тесты по темам лекций;
- компьютерные тесты по темам лабораторных работ;
- проверка выполнения лабораторных работ и защита индивидуальных заданий при выполнении лабораторных работ;
- компьютерное итоговое тестирование.

Методика формирования итоговой оценки

Итоговая оценка выставляется в соответствии с правилами проведения аттестации постановления №53 от 29 мая 2012 г., положения о рейтинговой системе БГУ(ред. 2015 г.), критериями оценки студентов (10 баллов).

[illegible]

Учебная программа пересмотрена и одобрена на заседании кафедры _____ (протокол № ____ от _____ 20 ____ г.)
(название кафедры)

В.В. Скакун
(И.О.Фамилия)

С.В. Малый
(И.О.Фамилия)